Refining Convolutional Neural Network Detection of Small-Bowel Obstruction in Conventional Radiography

Phillip M. Cheng[1]
Khoa N. Tran
Gilbert Whang
Tapas K. Tejura

OBJECTIVE. The purpose of this study was to evaluate improvement of convolutional neural network detection of high-grade small-bowel obstruction on conventional radiographs with increased training set size.

MATERIALS AND METHODS. A set of 2210 abdominal radiographs from one institution (image set 1) had been previously classified into obstructive and nonobstructive categories by consensus judgments of three abdominal radiologists. The images were used to fine-tune an initial convolutional neural network classifier (stage 1). An additional set of 13,935 clinical images from the same institution was reduced to 5558 radiographs (image set 2) primarily by retaining only images classified positive for bowel obstruction by the initial classifier. These images were classified into obstructive and nonobstructive categories by an abdominal radiologist. The combined 7768 radiographs were used to train additional classifiers (stage 2 training). The best classifiers from stage 1 and stage 2 training were evaluated on a held-out test set of 1453 abdominal radiographs from image set 1.

RESULTS. The ROC AUC for the neural network trained on image set 1 was 0.803; after stage 2, the ROC AUC of the best model was 0.971. By use of an operating point based on maximizing the validation set Youden *J* index, the stage 2–trained model had a test set sensitivity of 91.4% and specificity of 91.9%. Classification performance increased with training set size, reaching a plateau with over 200 positive training examples.

CONCLUSION. Accuracy of detection of high-grade small-bowel obstruction with a convolutional neural network improves significantly with the number of positive training radiographs.

Abdominal radiography is an inexpensive and commonly available screening test for evaluating for small-bowel obstruction. However, accurate discrimination of small-bowel obstruction on radiographs can be challenging, particularly for hospitalized patients in whom differentiation from ileus is required [1]. Furthermore, although detection accuracy for small-bowel obstruction has been reported to correlate with radiologist experience [2], immediate access to expert interpretation may not be readily available in all clinical settings. As a result, computer-assisted decision support for analyzing radiographs for small-bowel obstruction may be useful to clinicians and radiology trainees and for prioritizing studies in a large radiography worklist.

A branch of machine learning called deep learning has been used to achieve breakthrough performance improvements in image classification [3, 4]. Unlike prior image analysis systems that required hand-engineering of image features by human experts, a deep learning system for image classification entails use of a multilayer artificial neural network to learn the image features best suited for a given classification task. Convolutional neural networks are a type of deep learning architecture particularly suited for image analysis. These networks have a hierarchy of layers in which lower levels process small grids of pixels within an image to detect the presence of simple features. Higher-level nodes use the output from the lower levels to assess the image for more complex features. Examples of recent applications of convolutional neural networks for analyzing clinical radiographs include pneumonia detection on chest radiographs [5, 6] and estimation of pediatric bone age [7, 8].

Deep learning systems for image classification typically require large numbers of labeled training examples to learn the features that discriminate different image types. This

Keywords: artificial neural networks, deep learning, digital image processing, machine learning, small-bowel obstruction

doi.org/10.2214/AJR.18.20362

Received July 8, 2018; accepted after revision August 13, 2018.

Supported in part by Nvidia Corporation, which donated the Titan Xp graphics processing unit used for this research.

[1] All authors: Department of Radiology, Keck School of Medicine of USC, 1441 Eastlake Ave, Ste 2315B, Los Angeles, CA 90033. Address correspondence to P. M. Cheng (Phillip.Cheng@med.usc.edu).

AJR 2019; 212:342–350

0361–803X/19/2122–342

data requirement is mitigated by transfer learning, whereby a system that has learned useful features with one kind of data, such as color photographs, may reuse those features to efficiently learn to classify a related kind of data, such as gray-scale radiographs. Transfer learning is particularly useful in the medical domain, in which labeled training data can be difficult to acquire. However, even transfer learning may be insufficient to overcome problems in obtaining sufficient positive training examples. This was shown in a recent study [9] of neural network detection of small-bowel obstruction on abdominal radiographs in which radiologists deemed only 74 of 3663 (2%) consecutive radiographs to be consistent with complete or high-grade small-bowel obstruction.

The purpose of this study was to evaluate the performance improvement in convolutional neural network detection of high-grade small-bowel obstruction on radiographs with an increasing number of positive training examples. In this study, the acquisition of additional positive training examples from the PACS was facilitated by screening images using a weak neural network classifier trained on a limited number of positive examples from the prior study [9].

Materials and Methods

Institutional review board approval was obtained for retrospective data collection and analysis in this study.

Image Acquisition

A dataset obtained from a prior pilot study [9] composed of 3663 supine abdominal radiographs of 1299 patients obtained from January to June 2016 at a single institution (image set 1) was used to train the initial neural network classifier. These images had been cropped along the long dimension to squares, resized to 512 × 512 pixel resolution, and classified independently and without clinical information by three fellowship-trained abdominal radiologists [9] into categories of “normal, equivocal, or low-grade partial small-bowel obstruction,” or “complete or high-grade partial small-bowel obstruction,” according to the classification proposed by Silva et al. [10]. The majority judgment among the three reviewers was used as the ground truth for neural network training and validation. The images were randomized into training, validation, and test sets, such that the patients in each set were disjoint (Fig. 1A). The training and validation sets were used to fine-tune an initial neural network classifier (see later, Neural Network Training and Testing).

Subsequently, a new dataset composed of 13,935 images from consecutive supine abdominal radiographic studies of 4103 patients performed from July 2016 to December 2017 (image set 2) was obtained from the PACS of the same institution as image set 1. Images were cropped along the longer dimension to squares, and images were excluded that the initial neural network classified as negative for high-grade small-bowel obstruction. The other 6429 images in this set were then reviewed at 512 × 512 pixel resolution without clinical information by an abdominal radiologist with 9 years of postfellowship clinical experience. Images of scanned documents and incorrect anatomy were excluded from the set, as were images deemed to be of insufficient quality for interpretation, leaving 5558 supine abdominal radiographs of 2502 patients.

The 5558 images were then classified by the radiologist for the presence of complete or high-grade partial small-bowel obstruction using the same categories used for image set 1 (Fig. 1B). These images from image set 2 were then combined with the training and validation sets from image set 1, and the combined set was randomized into training and validation sets with disjoint groups of patients. These sets were then used to train several neural networks in a second round of training (see later, Neural Network Training and Testing). Final testing was performed with the held-out test set from image set 1, which was not used for evaluation during training of any of the neural network models (Fig. 1C).

Neural Network Training and Testing

Images were down-sampled by bicubic interpolation to a resolution of 299 × 299 pixels to match the input layers of the neural networks. For initial (stage 1) training on image set 1 (Fig. 1A), fine-tuning was performed on the Inception-v3 convolutional neural network [11] with weights pretrained on the ImageNet Large Scale Visual Recognition Challenge [12]. The weights of the neural network were fixed except for a final fully connected layer, which was retrained with the training subset of image set 1. Attempts at end-to-end training of the network with the small number of positive examples in image set 1 resulted in overfitting and poor test performance (data not shown).

Data augmentation was performed during training with random rotation of images up to 40°, horizontal flipping, horizontal and vertical shifting up to 20%, and magnification up to 20%. All neural network training and testing was performed at a workstation with a 3.07-GHz processor (Xeon ×5675, Intel), 16 GB RAM, and a 12-GB graphics card (Titan Xp, Nvidia). Training was performed with the Keras 2.1.5 neural network library with TensorFlow 1.7 as the backend tensor manipulation framework [13, 14]. For training, an Adam optimizer was used with standard parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$) [15] at a learning rate of 10^{-4} (logarithmic scale) over 90 epochs with a batch size of 8.

Each epoch represents one complete presentation of the training set to the model. Training cross-entropy loss values were weighted by a factor of 10 for positive training examples of small-bowel obstruction, to compensate for the numeric imbalance of positive versus negative examples in the training set (higher weighting factors did not appear to improve model performance). Validation cross-entropy loss was calculated after each epoch, and the model with the lowest validation loss (achieved at 43 epochs) was saved. A ROC curve was constructed by varying the threshold for diagnosis of high-grade small-bowel obstruction and evaluating the performance of the model against the validation set. The threshold corresponding to the maximum Youden *J* index (defined as sensitivity + specificity – 1) was used to apply the model to filter image set 2 (Fig. 1B).

A second round (stage 2) of training was subsequently performed on a combined dataset consisting of the filtered image set 2 and the training and validation sets of image set 1 (Fig. 1C). Tested architectures were Inception-v3, Inception-ResNet V2 [16], ResNet50 [17], Xception [18], DenseNet-121 [19], DenseNet-169 [19], and DenseNet-201 [19]. Training was performed end to end from weights pretrained on the ImageNet Large Scale Visual Recognition Challenge. Data augmentation was again performed during training with random rotation of images up to 40°, horizontal flipping, horizontal and vertical shifting up to 20%, and magnification up to 20%. Training was performed with Keras 2.1.5 with an Adam optimizer with standard parameters at a learning rate of 10^{-4} up to 150 epochs with a batch size of 8. As in stage 1, training cross-entropy loss values were weighted by a factor of 10 for positive examples of small-bowel obstruction to compensate for the numeric imbalance of positive versus negative examples in the training set. Validation cross-entropy loss was calculated after each epoch, and the model with the lowest validation loss was saved. For the model with the lowest validation loss of each architecture, ROC AUCs for the validation set were calculated. The model with the highest validation set ROC AUC was used for final testing.

The final trained network was evaluated on the held-out test set images from image set 1 (Fig. 1C). These images had not been used for either stage 1 or stage 2 training.

Calculation of all 95% CIs of the ROC AUC for validation or test set performance was performed

TABLE 1: Clinical Indications for Abdominal Radiographic Studies Reviewed in Filtered Image Set 2

Indication	No. of Studies
Tube placement	1827 (38.3)
Pain or distention	1026 (21.5)
Ileus or obstruction	741 (15.5)
Foreign body	248 (5.2)
Nausea or vomiting	222 (4.7)
Stones	171 (3.6)
Constipation	140 (2.9)
Postoperative	97 (2.0)
Perforation	52 (1.1)
Other	251 (5.3)

Note—Values in parentheses are percentages.

on 100,000 bootstrap samples sampled with replacement from the validation or test set.

A saliency map highlights pixels that are most important for the classification of a particular image by a neural network. It is calculated from the gradient of the class score with respect to the image pixels [20]. Saliency maps were created from test set images classified by the stage 1– and stage 2–trained Inception-v3 networks by use of the keras-vis visualization toolkit [21].

Results

Image set 1 consisted of 3663 radiographs from 3027 studies of 1299 patients (743 men, 556 women; mean age, 59.1 ± 16.4 [SD] years; range, 18–97 years). Clinical indications for the studies in this dataset have been previously reported [9]. The images in image set 2 used for stage 2 training of the neural network model consisted of 5558 radiographs from 4775 studies of 2502 patients (1547 men, 955 women; mean age, 59.9 ± 15.9 years; range, 17–101 years). As with image set 1, the most common clinical indications for the studies were assessment of tube placement, pain or distention, and ileus or obstruction (Table 1).

After stage 1 training on Inception-v3, the validation set ROC AUC for the model with the lowest validation loss was 0.887 (Fig. 2A). The operating point with the highest Youden *J* index corresponded to sensitivity of 100% and specificity of 62.2% for the stage 1 validation set.

After stage 2 training, the model architecture with the largest validation set ROC AUC was Inception-v3. However, there was substantial overlap of the 95% CIs of the ROC AUC for all trained architectures (Table 2).

TABLE 2: ROC AUC for Stage 2 Validation Set Classification Performance

Architecture	ROC AUC
Inception-v3 [11]	0.980 (0.971–0.989)
Inception ResNet V2 [16]	0.979 (0.968–0.987)
DenseNet 201 [19]	0.978 (0.967–0.987)
DenseNet 169 [19]	0.970 (0.953–0.983)
DenseNet 121 [19]	0.969 (0.953–0.982)
Xception [18]	0.969 (0.954–0.981)
ResNet 50 [17]	0.959 (0.933–0.979)

Note—Numbers in square brackets are references. Values in parentheses are 95% CIs calculated from classification of 100,000 bootstrap samples of the validation set sampled with replacement from the validation set.

The Inception-v3 model was used for evaluation on the final test set.

Test set performance significantly improved after stage 2 training. The ROC AUC for Inception-v3 after stage 1 training was 0.803 (95% CI, 0.744–0.858) (Fig. 2B). After stage 2 training, the ROC AUC for Inception-v3 was 0.971 (95% CI, 0.951–0.986). Operating points were selected for the stage 1– and stage 2–trained Inception-v3 models on the basis of the highest Youden *J* index for the corresponding stage 1 or stage 2 validation set. For the stage 1–trained model, this corresponded to sensitivity of 82.9% and specificity of 63.2% for the test set (Fig. 2C). For the stage 2–trained model, sensitivity was 91.4% and specificity was 91.9% for the test set (Fig. 2D).

For each of the three false-negative test set images for the stage 2–trained model, one of the three judging radiologists had also considered the image negative for high-grade small-bowel obstruction. Of the 115 false-positive test set images in the stage 2–trained model, 45 images had been considered positive for high-grade small-bowel obstruction by one of the three judging radiologists. The other 70 images were reviewed by a separate abdominal radiologist with 9 years of post-fellowship experience. Of these, 56 (80%) appeared consistent with ileus or at least low-grade small-bowel obstruction.

Saliency maps highlight the most important pixels for a neural network in classifying a particular image. Saliency maps for the stage 2–trained model of positive test set images were found to highlight more localized groups of pixels corresponding to dilated segments of bowel compared with saliency maps generated with the stage 1–trained model (Fig. 3).

Five incremental training sets were constructed to evaluate the effect of training set size on test set performance of the neural network. These training sets started with the images from image set 1 that had been randomized into the stage 2 training set and were formed by incrementally adding 0%, 20%, 40%, 60%, and 80% of the positive training examples from image set 2 (Fig. 4A). A proportionate number of negative training examples were included in each incremental training set. A constant validation set equivalent to that used in stage 2 training was used. Training with these incremental image sets was performed with the same Inception-v3 model architecture and technique hyperparameters as stage 2 training. For each incremental training set, the model with the lowest validation set ROC AUC was evaluated with the stage 1 test set. The test set ROC AUC increased as a function of the number of positive examples in the training set but appeared to plateau above 200 positive training examples (Fig. 4B).

In an evaluation schema adapted from that performed on the CheXNet neural network trained for pneumonia detection [5], we compared the performance of the stage 2–trained Inception-v3 network with the performance of the three radiologists who classified the image test set. Specifically, for each image in the test set, there are three labels from the radiologists and one label from the neural network. We computed sensitivity, specificity, positive predictive value, negative predictive value, and F1 score (harmonic mean of sensitivity and positive predictive value) for each of the three radiologists and for the neural network, using each of the other three labels as ground truth. The mean of the resulting score values was calculated, and 95% CIs were calculated with score values from classification of 100,000 bootstrap samples sampled with replacement from the test set. The neural network had low positive predictive value and slightly higher negative predictive value than the radiologists. The 95% CIs for F1 score exhibited considerable overlap between the radiologists and neural network (Table 3).

Discussion

Interest in the application of deep learning to problems in radiology has been tempered by questions regarding the amount of labeled clinical data required for training an effective deep learning system. Because

TABLE 3: Test Set Classification Performance of Radiologists Versus Stage 2–Trained Inception-v3 Model

Reader	Sensitivity (%)	Specificity (%)	Positive Predictive Value	Negative Predictive Value	F1
Radiologist 1	28.5 (19.4–37.6)	99.6 (99.3–99.8)	0.782 (0.667–0.884)	0.950 (0.942–0.959)	0.392 (0.288–0.483)
Radiologist 2	38.0 (28.0–47.5)	99.1 (98.7–99.5)	0.658 (0.546–0.765)	0.956 (0.948–0.963)	0.423 (0.325–0.507)
Radiologist 3	65.5 (55.1–74.6)	96.4 (95.5–97.2)	0.431 (0.364–0.498)	0.977 (0.972–0.982)	0.449 (0.372–0.519)
Mean of radiologists	44.0 (36.9–50.6)	98.4 (98.0–98.7)	0.623 (0.545–0.694)	0.961 (0.955–0.967)	0.421 (0.336–0.497)
Inception-v3	82.9 (74.3–90.2)	92.5 (91.1–93.7)	0.279 (0.227–0.333)	0.993 (0.990–0.995)	0.396 (0.329–0.458)

Note—For each reviewer, the judgment of each of the other three reviewers (including the neural network) was used as ground truth, and mean statistics were calculated across the other three reviewers. F1 = harmonic mean of sensitivity and positive predictive value. Values in parentheses are 95% CIs calculated from classification of 100,000 bootstrap samples of the test set sampled with replacement from the test set.

of privacy concerns that limit data sharing, large medical imaging datasets are uncommon; moreover, concise expert labeling of these datasets can be laborious and expensive. Transfer learning from large photographic image collections, such as ImageNet [12], can be used to reduce training set requirements but may not be sufficient to produce adequate training performance in very limited clinical datasets.

In this work, we sought additional training data to improve classification specificity of a convolutional neural network for detecting high-grade small-bowel obstruction on abdominal radiographs. Transfer learning applied to the small number of positive abdominal radiographs in image set 1 resulted in a system with low specificity. Prior work has shown that augmentation of these limited training abdominal radiographs by image transformations that do not affect the classification label (rotation, mirroring, and scaling) did not improve test set performance, likely because the images did not include sufficient variety in the possible radiographic patterns of small-bowel obstruction [9].

Assuming an incidence of high-grade small-bowel obstruction in a set of abdominal radiographs of 2% (the approximate incidence in image set 1), 5000 images would have to be reviewed to find an additional 100 positive training examples. One approach to reducing the expert human labor required for image review would be to scan the associated radiology reports for statements affirming the presence of high-grade small-bowel obstruction; however, this would have required either human review or training of a natural language classifier. In this study, we instead used the weak classifier trained on image set 1 to enrich the percentage of positive examples in a new set of abdominal radiographs (image set 2). The radiologist reviewing filtered image set 2 judged 462 radiographs positive for high-grade small-bowel obstruction in a set of 5558 images (8.3%).

The test set performance of the stage 2–trained network significantly increased compared with the stage 1 network in terms of both ROC AUC and sensitivity and specificity at an operating point based on validation set performance. The saliency maps for the stage 2–trained network also appeared more comprehensible than those for the stage 1–trained network, more consistently identifying dilated segments of small bowel as important to the classification of small-bowel obstruction. This recognition of dilated bowel segments as a feature of small-bowel obstruction was never explicitly taught to the neural network but instead emerged from the repeated exposure of the neural network to positive training examples.

Using incremental training sets, we found a gradual increase in test set performance with training set size, until the number of positive examples in the training set exceeded 200 images. This finding may have implications for training set requirements for other clinical detection tasks in conventional radiography, though we expect that the threshold level of positive training images required to reach a performance plateau will vary according to the complexity of the classification task. For instance, slight improvements in pediatric bone age assessment by a convolutional neural network were seen with up to 12,611 training images [8].

Because the ground truth for the test set was based on consensus radiologist judgments rather than clinical or imaging follow-up, unbiased comparison of the trained neural network with human radiologists was not possible in this study. However, we adapted a previously proposed multiway comparison scheme to evaluate the relative performance of each classifier (human or neural network) using the others as ground truth labels. Under this scheme, test set sensitivity for the neural network was higher and specificity was lower than those of the human radiologists.

The overall balanced sensitivity and specificity of a neural network in the test set are related to the arbitrary choice of using as the operating point the maximum Youden *J* index in the validation set, weighing sensitivity and specificity equally. However, the low prevalence of high-grade small-bowel obstruction in the test set, which was formed from a consecutive retrospective series of clinical studies, resulted in a much lower positive predictive value for the neural network and slightly higher negative predictive value than for the radiologists. Of note, the calculated F1 scores of the radiologists and neural network were comparable, having overlapping 95% CIs. The similarities in F1 scores and marked differences in positive predictive values among the classifiers indicate that the F1 score should not be used alone to summarize classification performance, even though high F1 scores have been used to claim performance superiority of a deep neural network compared with radiologists for detecting pneumonia on chest radiographs [5].

Despite the low positive predictive value of the neural network in the test set, its high sensitivity and negative predictive value at the selected operating point may be desirable in a clinical setting for a screening test such as abdominal radiography. Many of the false-positive results in the set appeared to be consistent with low-grade small-bowel obstruction or ileus. We suspect that neural network specificity in the test set might have improved had the radiologists who provided the original judgments on image set 1 also provided the labels for the additional training images from image set 2. It does not appear that substantial increases in training set size would necessarily improve classification performance. Furthermore, classification performance in stage 2 was not significantly different among a variety of recent deep neural network architectures.

Limitations

There were several limitations to this study. First, Because the radiologist providing judgments on image set 2 was not one of the radiologists providing judgments on image set 1, the degree of enrichment of image set 2 with positive examples might have been overestimated because of increased radiologist sensitivity in interpreting high-grade obstruction. However, the improvements in both sensitivity and specificity in the image set 1 test set for the neural network after stage 2 training indicate similarity of the radiologist judgments for image sets 1 and 2.

Second, the number of positive examples in the test set was small. Using a trained neural network to prescreen clinical radiographs to facilitate expansion of positive test set examples would clearly introduce bias at test time in the detection of positive examples by the same neural network. As a result, considerable human expert time and labor are still required to increase the number of positive examples in the test set in an unbiased manner representative of the patient population imaged.

Third, ground truth was determined by radiologist judgments rather than clinical outcomes, which would have been difficult to determine consistently given the large number of radiographs with varying degrees of clinical follow-up. The goal of this study was to evaluate whether deep neural networks can be used for assessments of high-grade small-bowel obstruction similar to those of human experts in an experimental context in which ancillary clinical information is not available. However, nonspecific bowel gas patterns that in certain clinical settings could raise suspicion for small-bowel obstruction, such as the so-called gasless abdomen [22], would have been classified in the negative-equivocal category by the radiologist judges of our training and test sets. Future radiographic datasets with ground truth labels verified by clinical outcomes may provide more accurate training data and performance measures for deep learning systems.

Fourth, the ground truth labels were simple binary assessments of the presence of high-grade small-bowel obstruction on supine radiographs. The binary assessment used in this study has been proposed as a primary distinction on abdominal radiographs to be made for further clinical management [1, 10]. However, other potentially useful clinical labels, such as low-grade small-bowel obstruction, ileus, and the presence of other important image features, such as pneumatosis, free intraperitoneal gas, and abdominal calcifications, were not assigned to the image sets in this study. At present, training for such distinctions would require extensive human expert labor in labeling the images. Our focus on supine abdominal radiographs arises from the observation that most of the abdominal radiographs in the PACS of our institution are obtained with the patient supine. Although upright radiographs may provide more specific image details on small-bowel obstruction, they were not incorporated into the training and test sets of this study.

Conclusion

The classification performance of a deep convolutional neural network in detecting high-grade small-bowel obstruction significantly improved with the number of positive training examples up to a threshold of at least 200 positive examples. Expansion of the training set was facilitated by screening of PACS images by use of a weak classifier trained on few positive training examples.

References

1. Paulson EK, Thompson WM. Review of small-bowel obstruction: the diagnosis and when to worry. *Radiology* 2015; 275:332–342
2. Thompson WM, Kilani RK, Smith BB, et al. Accuracy of abdominal radiography in acute small-bowel obstruction: does reviewer experience matter? *AJR* 2007; 188:[web]W233–W238
3. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521:436–444
4. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. *RadioGraphics* 2017; 37:2113–2131
5. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv website. arxiv.org/abs/1711.05225. Last revised December 25, 2017. Accessed May 19, 2018
6. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017; 284:574–582
7. Kim JR, Shim WH, Yoon HM, et al. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. *AJR* 2017; 209:1374–1380
8. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology* 2018; 287:313–322
9. Cheng PM, Tejura TK, Tran KN, Whang G. Detection of high-grade small bowel obstruction on conventional radiography with convolutional neural networks. *Abdom Radiol (NY)* 2018; 43:1120–1127
10. Silva AC, Pimenta M, Guimarães LS. Small bowel obstruction: what to look for. *RadioGraphics* 2009; 29:423–439
11. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. arXiv website. arxiv.org/abs/1512.00567. Last revised December 11, 2015. Accessed May 14, 2017
12. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015; 115:211–252
13. Keras website. Keras: the Python deep learning library. keras.io. Accessed April 18, 2018
14. Adabi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems. download.tensorflow.org/paper/whitepaper2015.pdf. Published November 9, 2015. Accessed April 18, 2018
15. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv website. arxiv.org/abs/1412.6980. Last revised December 22, 2014. Accessed April 19, 2018
16. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. arXiv website. arxiv.org/abs/1602.07261. Last revised August 23, 2016. Accessed May 18, 2018
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv website. arxiv.org/abs/1512.03385. Published December 10, 2015. Accessed May 18, 2018
18. Chollet F. Xception: deep learning with depthwise separable convolutions. arXiv website. arxiv.org/abs/1610.02357. Last revised April 4, 2017. Accessed May 18, 2018
19. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. arXiv website. arxiv.org/abs/1608.06993. Last revised January 28, 2018. Accessed May 18, 2018
20. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv website. arxiv.org/abs/1312.6034. Last revised April 19, 2014. Accessed July 24, 2017
21. Kotikalapudi R. Keras visualization toolkit. raghakot.github.io/keras-vis/. 2017. Accessed May 18, 2018
22. Thompson WM. Gasless abdomen in the adult: what does it mean? *AJR* 2008; 191:1093–1099

(Figures start on next page)

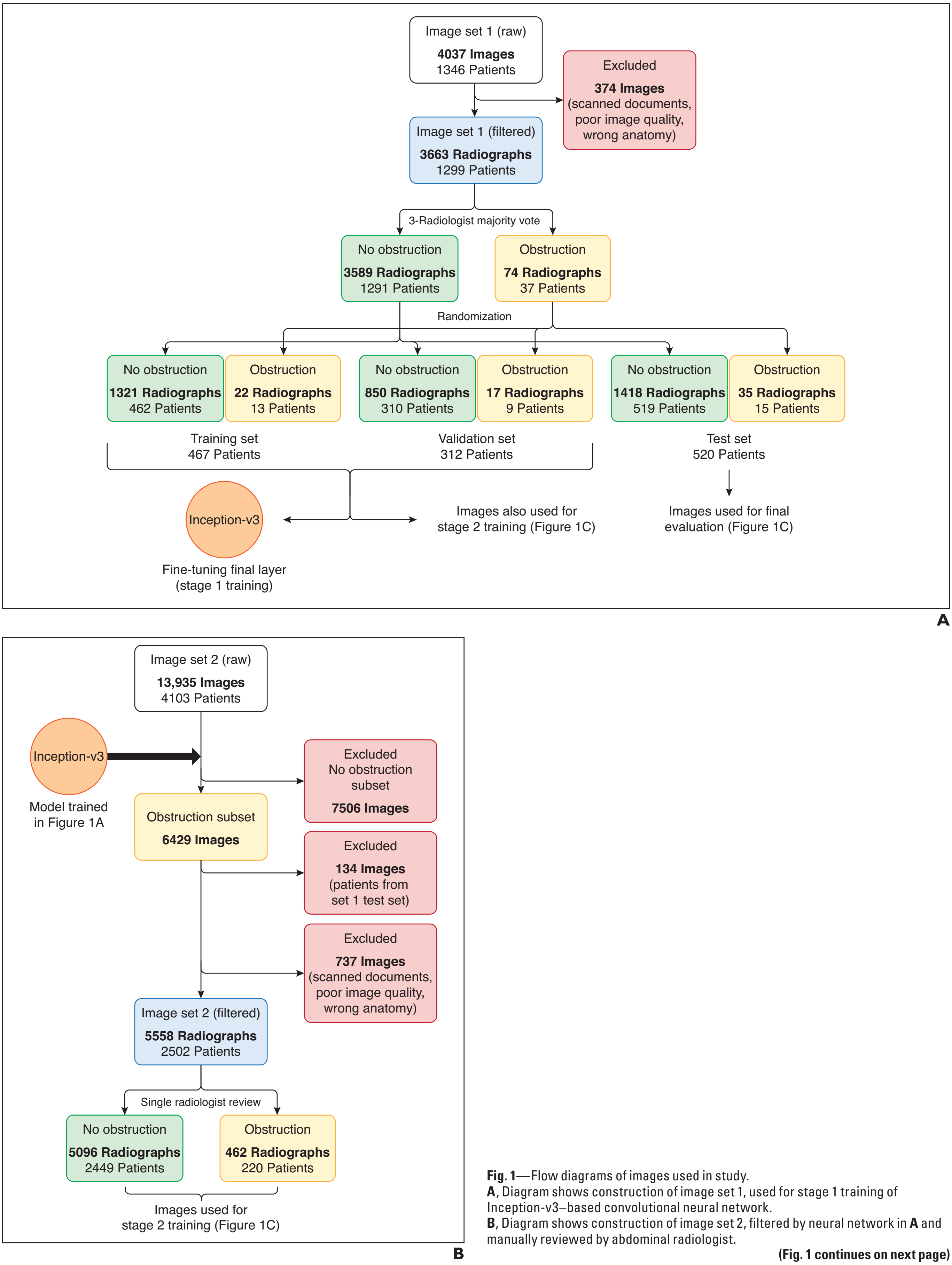

Fig. 1—Flow diagrams of images used in study.
A, Diagram shows construction of image set 1, used for stage 1 training of Inception-v3–based convolutional neural network.
B, Diagram shows construction of image set 2, filtered by neural network in **A** and manually reviewed by abdominal radiologist.

(Fig. 1 continues on next page)

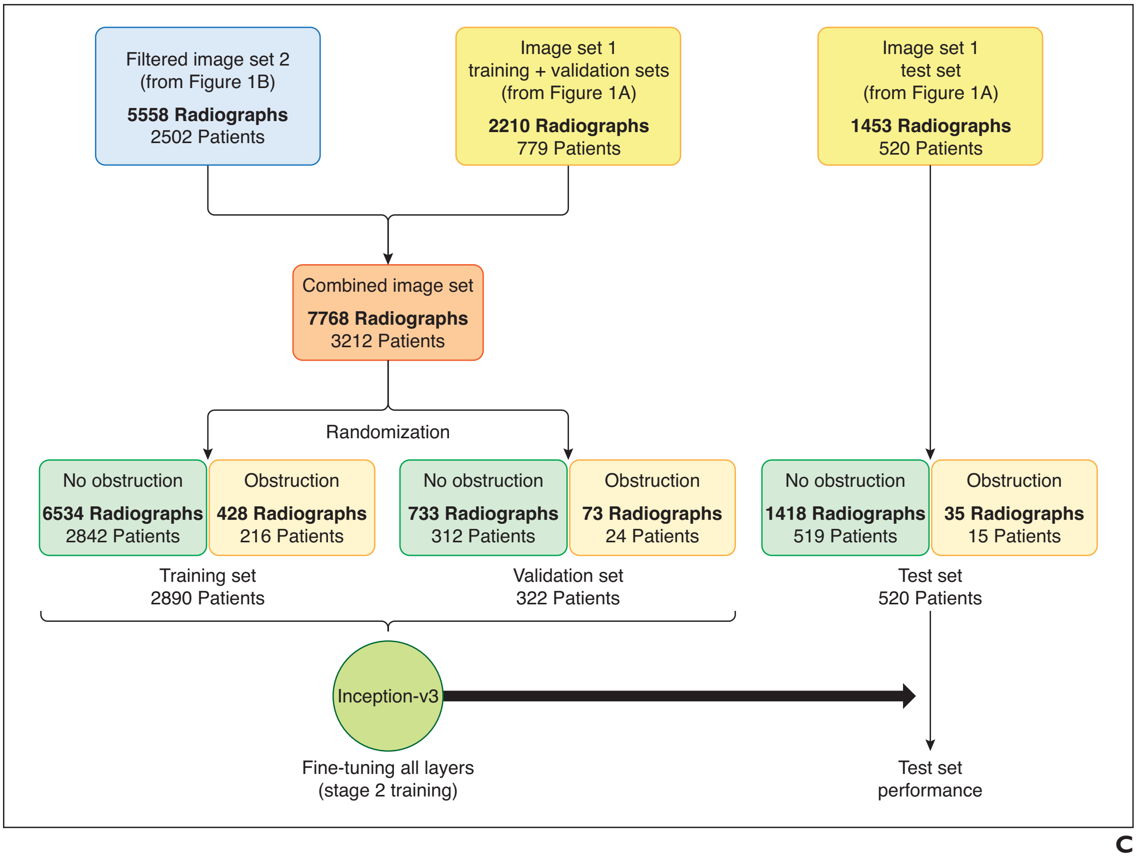

Fig. 1 (continued)—Flow diagrams of images used in study. **C**, Diagram shows combination of image sets 1 and 2, used for stage 2 training of Inception-v3–based convolutional neural network.

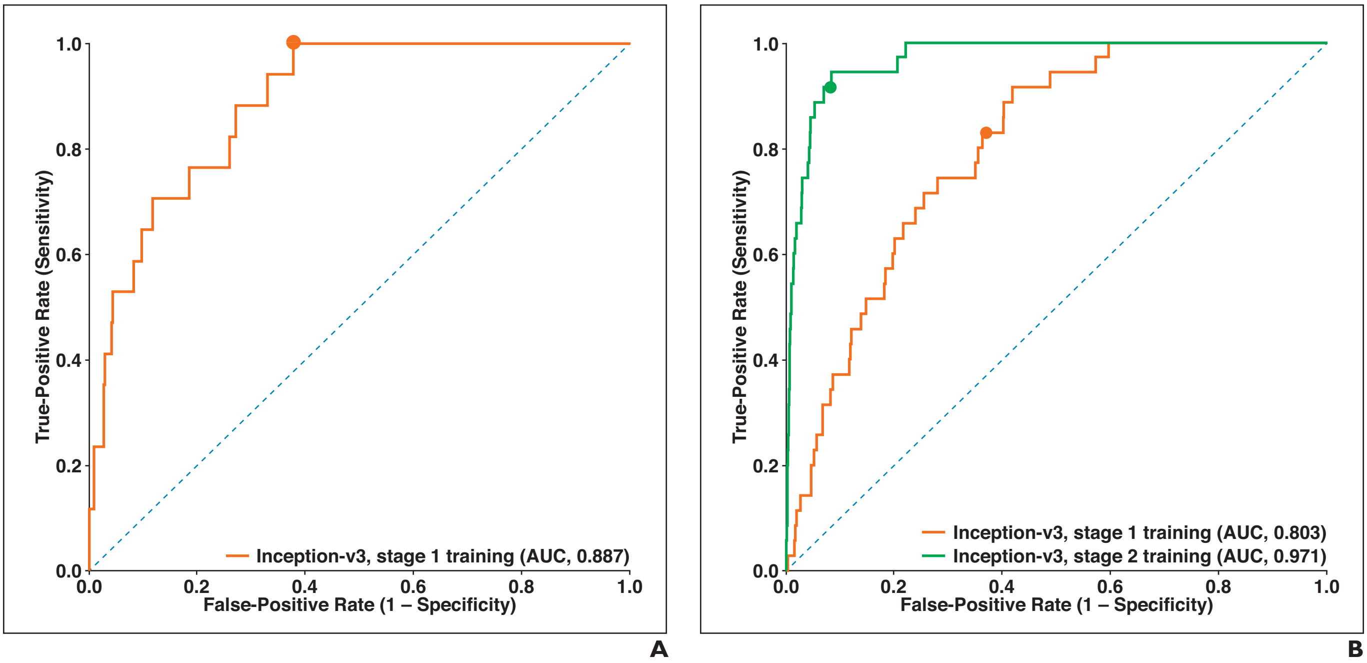

Fig. 2—Performance of convolutional neural networks for detecting small-bowel obstruction.
A, Graph shows ROC curve for validation set performance of Inception-v3 neural network after stage 1 training. Operating point with highest Youden *J* index (*circle*) was used for filtering image set 2.
B, Graph shows ROC curves for test set performance of trained Inception-v3 neural networks after stage 1 (*orange*) and stage 2 (*green*) training. Circles on curves correspond to operating points with highest Youden *J* indexes for validation sets.

(Fig. 2 continues on next page)

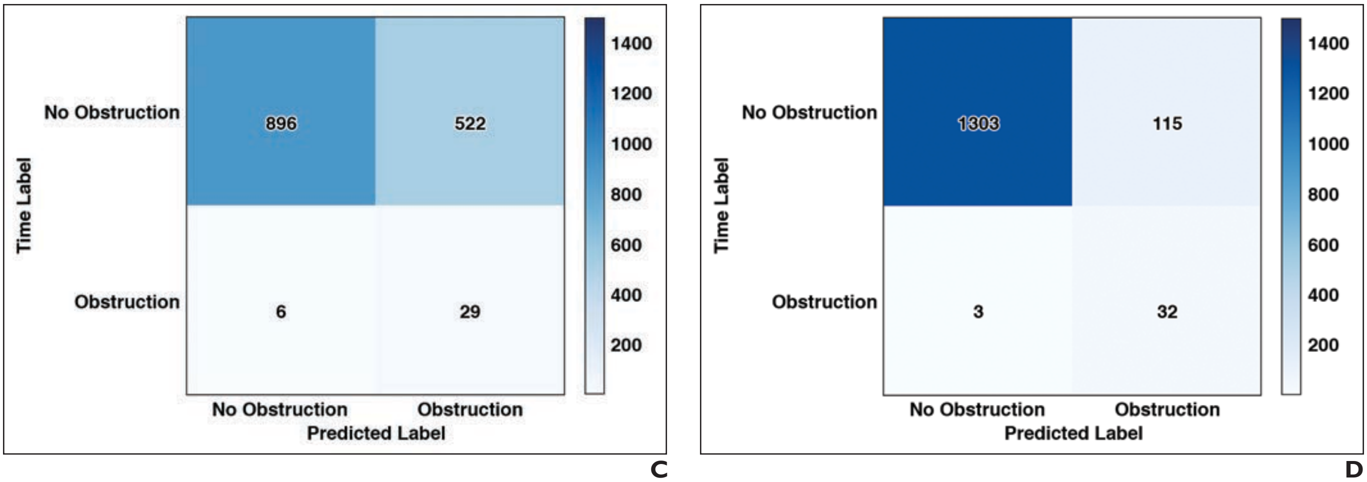

Fig. 2 (continued)—Performance of convolutional neural networks for detecting small-bowel obstruction.
C and **D**, Charts show test set confusion matrices for stage 1–trained (**C**) and stage 2–trained (**D**) neural networks based on operating points in **B**.

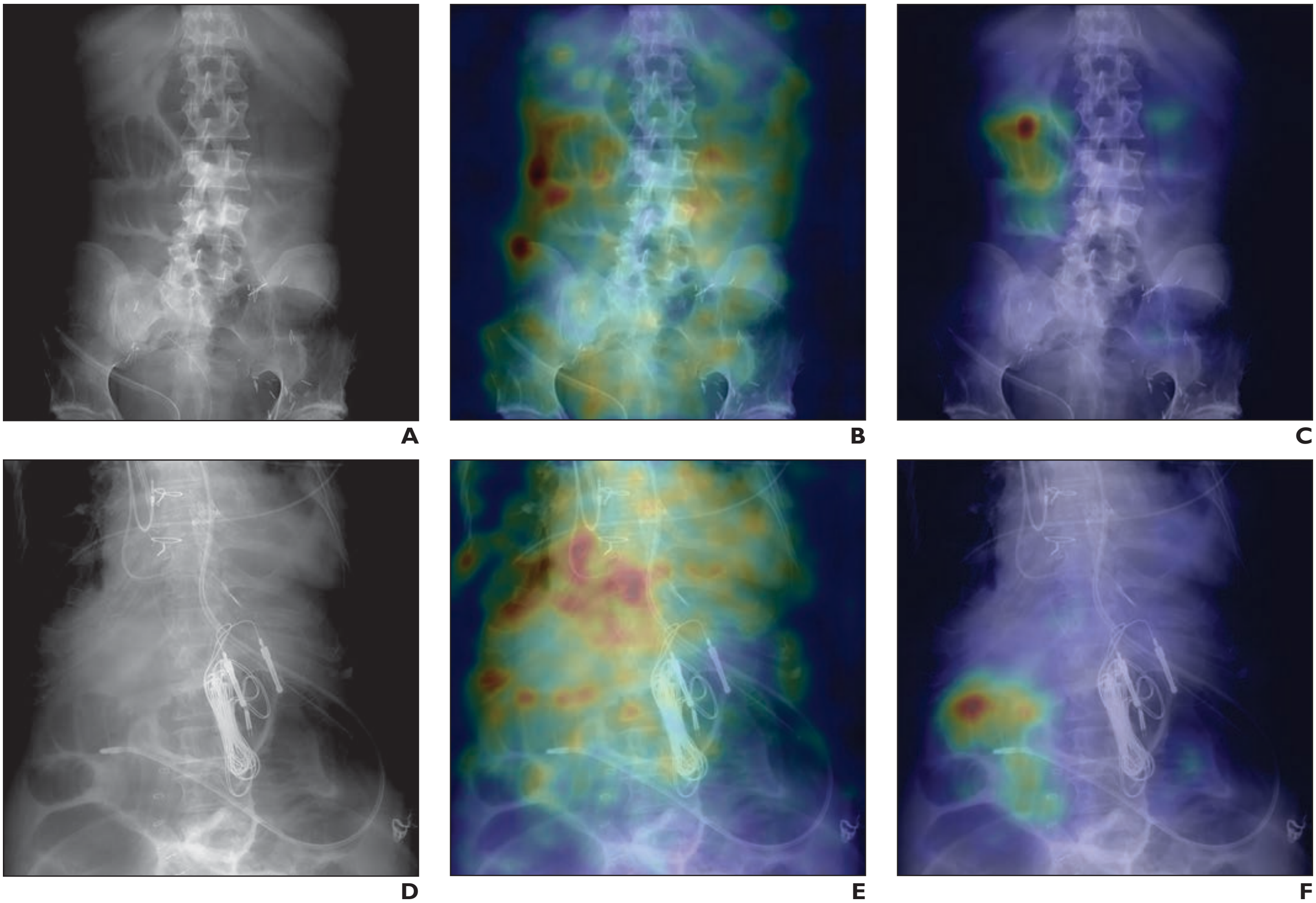

Fig. 3—Saliency maps of most important image pixels for classification. **A** and **D** were judged consistent with small-bowel obstruction by majority of radiologist reviewers and by both stage 1– and stage 2–trained Inception-v3 neural networks. However, stage 1–trained network highlights diffusely distributed important pixels in **B** and **E**, not corresponding to bowel gas pattern. Stage 2–trained network shows localization of important pixels in regions of dilated small bowel in **C** and **F**.
A–C, 60-year-old man with nausea and vomiting. Saliency maps for classifying radiograph in **A** show results from stage 1–trained network (**B**) and stage 2–trained network (**C**).
D–F, 77-year-old woman undergoing radiography for enteric tube placement. Saliency maps for classifying radiograph in **D** show results from stage 1–trained network (**E**) and stage 2–trained network (**F**).

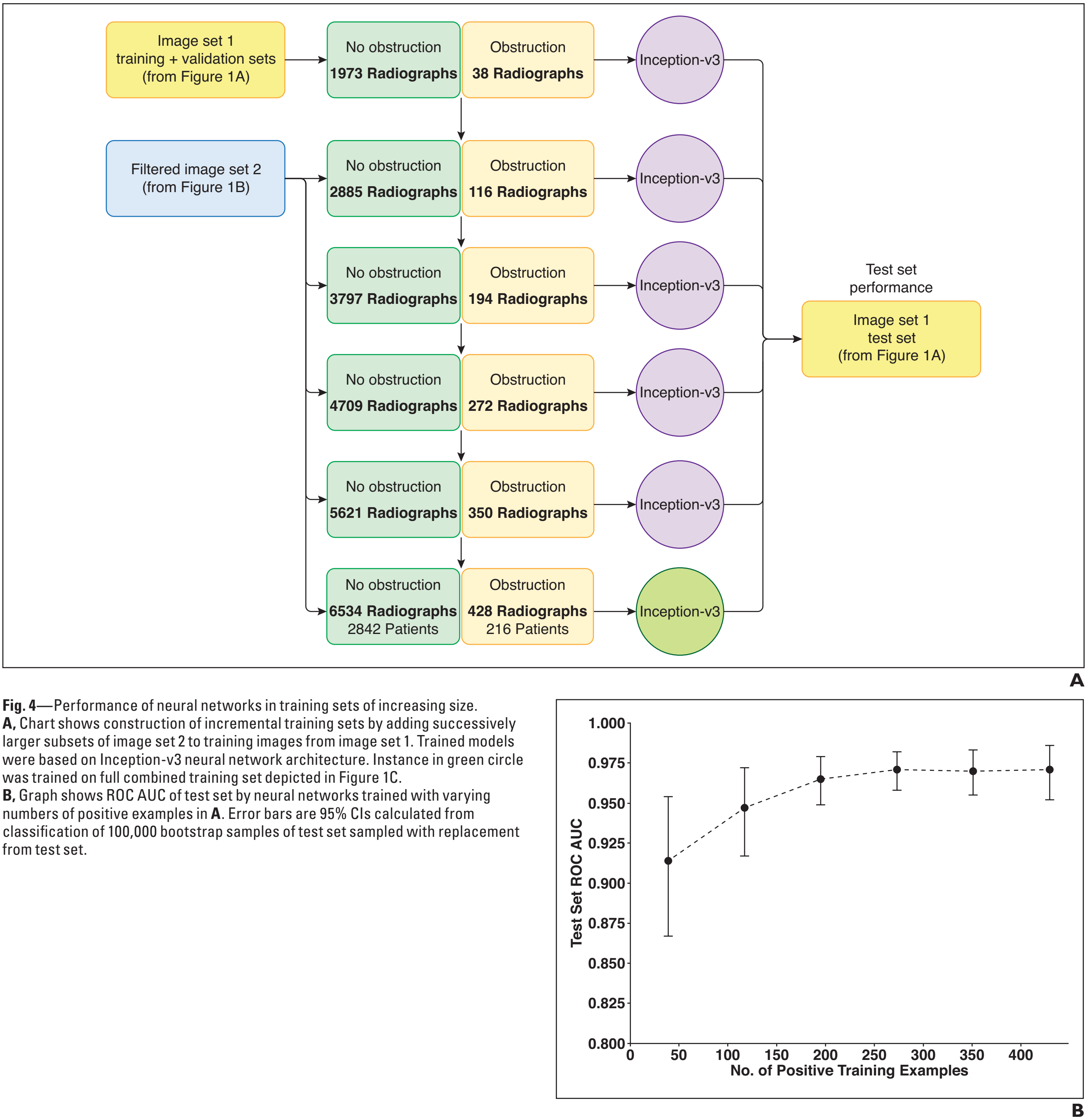

Fig. 4—Performance of neural networks in training sets of increasing size.
A, Chart shows construction of incremental training sets by adding successively larger subsets of image set 2 to training images from image set 1. Trained models were based on Inception-v3 neural network architecture. Instance in green circle was trained on full combined training set depicted in Figure 1C.
B, Graph shows ROC AUC of test set by neural networks trained with varying numbers of positive examples in **A**. Error bars are 95% CIs calculated from classification of 100,000 bootstrap samples of test set sampled with replacement from test set.